

A Systematic Review of Deepfake Detection Techniques: Cross-Dataset Generalization, Computational Efficiency, And Performance Evaluation

Nandini Kalshetty¹, Samiksha Kharosekar², Shreya Rokade³, Mrs. Punam Patil⁴

^{1,2,3,4}MIT Arts Commerce and Science College, Alandi

doi.org/10.64643/ijirt.209546-459

Abstract—Deepfake technology has emerged as a major challenge to the authenticity and trustworthiness of digital media. The rapid development of artificial intelligence and deep learning has made it possible to generate highly realistic manipulated images, videos, and other forms of digital content. As a result, reliable deepfake detection systems are increasingly important for digital forensics, cybersecurity, social media platforms, and the protection of information integrity. This paper presents a systematic review of deepfake detection techniques published between 2018 and 2025, with particular attention to deep learning, traditional machine learning, and classical image-processing approaches.

The review focuses on detection performance, computational efficiency, and cross-dataset generalization. Studies evaluated using FaceForensics++, the DeepFake Detection Challenge (DFDC), and Celeb-DF indicate that Transformer-based methods generally demonstrate stronger cross-dataset generalization than evaluated CNN-based approaches. However, this improved generalization comes with substantially higher computational requirements.

Traditional machine learning approaches, particularly Random Forest, can provide competitive performance in some settings while requiring fewer computational resources. Overall, performance degradation of approximately 10–15% across datasets remain a significant challenge. These findings indicate that future deepfake detection research should prioritize robustness, generalization, computational efficiency, and real-world deployment rather than focusing exclusively on accuracy.

Index Terms—Deepfake; Deepfake Detection; Machine Learning; Deep Learning; Computer Vision; Cross Dataset Generalization; Digital Forensics; Artificial Intelligence

I. INTRODUCTION

The rapid advancement of artificial intelligence has transformed the way digital content is created and distributed. One of the most significant developments in this area is deepfake technology, which uses machine learning and deep learning techniques to create or manipulate realistic digital media. Deepfakes can alter facial expressions, replace identities, synthesize speech, or generate completely artificial content that appears authentic. Although these technologies have legitimate applications in entertainment, education, and creative industries, their misuse creates serious concerns regarding misinformation, fraud, privacy, identity manipulation, and digital trust. The increasing realism of synthetic media has created a corresponding need for effective deepfake detection systems. Researchers have developed a wide variety of detection methods using deep learning, traditional machine learning, and classical image-processing techniques.

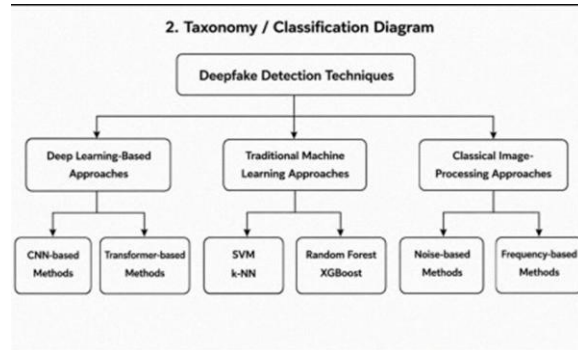
However, achieving high accuracy on a single dataset does not necessarily mean that a detector will perform reliably in real-world situations. A major issue identified in the reviewed literature is cross-dataset generalization. A model trained on one dataset may learn characteristics that are specific to that dataset, such as compression patterns, encoding properties, or manipulation artifacts, instead of learning features that are common to deepfakes in general.

The systematic review considered research published between 2018 and 2025 and examined studies using three widely used benchmark datasets: FaceForensics++, the DeepFake Detection Challenge (DFDC), and Celeb-DF. The review compares

different detection approaches according to their performance, computational requirements, and ability to generalize to previously unseen datasets.

II. DEEPFAKE DETECTION TECHNIQUES

Deepfake detection approaches can broadly be divided into three categories: deep learning techniques, traditional machine learning methods, and classical image-processing approaches.



2.1. Deep Learning-Based Approaches

Deep learning has become one of the most widely used approaches for detecting manipulated media. Convolutional Neural Networks (CNNs) are particularly important because they can automatically learn visual features from images and video frames. CNN-based detectors can identify inconsistencies in facial regions, textures, boundaries, and other visual characteristics associated with manipulation.

More recently, Transformer-based architectures have been investigated for deepfake detection. Transformers can model relationships between different parts of an input and may capture more generalized patterns than conventional CNN architectures.

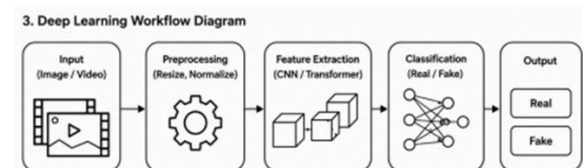
According to the reviewed findings, Transformer-based approaches showed better cross-dataset generalization than the evaluated CNN-based approaches. Their observed performance decline was approximately 11.33%, compared with declines exceeding 15% for CNN-based methods.

Despite their advantages, Transformer-based systems have an important limitation: computational cost. The reviewed approaches required approximately 3–5 times greater computational resources than the corresponding alternatives. This creates challenges for organizations or applications operating with limited computing resources.

2.2. Traditional Machine Learning Approaches

Traditional machine learning techniques remain relevant because they can offer a useful balance between computational efficiency and detection performance. These approaches generally, depend on extracted features that are subsequently classified using algorithms such as Random Forest and other conventional classifiers.

The review indicates that Random Forest methods demonstrated competitive performance in certain evaluation settings while requiring substantially fewer computational resources. This makes traditional machine learning particularly attractive for resource-constrained deployment.



2.3. Classical Image-Processing Techniques

Classical image-processing methods represent another category of deepfake detection. These techniques can analyze visual characteristics such as image artifacts, inconsistencies, texture abnormalities, and other measurable properties of manipulated media. Although these methods may be less computationally demanding than modern deep learning architectures, their effectiveness can be affected by the increasing sophistication of deepfake generation techniques.

III. DATASETS AND CROSS-DATASET GENERALIZATION

Evaluation datasets play an important role in determining the reliability of deepfake detection systems.

The reviewed studies considered three commonly used benchmark datasets: FaceForensics++, DFDC, and Celeb-DF. These datasets provide different examples of manipulated media and allow researchers to evaluate the effectiveness of detection methods under different conditions.

One of the most important findings of the review is that models often perform significantly better when tested under conditions similar to their training data. When models are evaluated on different datasets, performance generally decreases. Across the reviewed

methods, performance degradation of approximately 10–15% was observed when models were evaluated across different datasets.

This problem is particularly important for real-world deployment. In practical environments, a detector cannot assume that all incoming media will have the same resolution, compression, encoding, source, or manipulation method as its training data. A reliable detection system must therefore identify previously unseen manipulations and remain effective under changing conditions.

One possible explanation for poor cross-dataset performance is that some detection systems learn dataset-specific characteristics rather than generalized manipulation features. The review specifically highlights compression artifacts as an example of characteristics that may influence model decisions.

IV. PERFORMANCE VERSUS COMPUTATIONAL EFFICIENCY

An important consideration in deepfake detection is the trade-off between detection performance and computational efficiency. A model with very high detection accuracy may not be practical if it requires excessive computational resources.

The reviewed findings demonstrate this trade-off clearly. Transformer-based architectures provided better cross-dataset generalization, but their computational requirements were approximately 3–5 times higher. Traditional machine learning approaches such as Random Forest, on the other hand, achieved competitive performance in some evaluation settings while requiring substantially fewer resources.

This suggests that there is no single detection method that is optimal for every application. High-resource environments may be able to use computationally expensive deep learning architectures, whereas resource-constrained environments may benefit from simpler machine learning techniques. Therefore, evaluation of deepfake detection systems should not be based only on accuracy. Computational cost, robustness, generalization, and deployment requirements should also be considered.

V. DISCUSSION

The findings of this systematic review demonstrate that deepfake detection remains an evolving research

problem. Although modern deep learning architectures have significantly improved detection capabilities, the challenge of generalizing to unseen datasets remains unresolved.

Transformer-based approaches appear promising because of their relatively better cross dataset performance. However, their increased computational requirements may limit their practical use. CNN-based approaches remain important but can experience larger performance reductions when transferred between datasets. Traditional machine learning approaches offer an alternative where computational efficiency is a major requirement. The results also demonstrate the importance of designing evaluation methodologies that more closely represent real-world conditions. A detector that performs well on a benchmark dataset may not necessarily be reliable when exposed to new manipulation techniques, different compression levels, or media from previously unseen sources.

Consequently, future research should move beyond accuracy-focused comparisons and consider multiple dimensions of system quality. These include cross-dataset generalization, computational efficiency, robustness to different manipulations, and practical deployment requirements.

VI. FUTURE SCOPE

Future research in deepfake detection should focus on developing models capable of identifying previously unseen manipulation techniques. Such systems should learn generalized characteristics of manipulated media rather than depending heavily on dataset specific artifacts.

Another important research direction is improving computational efficiency. Although Transformer-based systems show promising generalization capabilities, their higher resource requirements create deployment challenges. Developing lightweight architectures or efficient hybrid approaches could help address this problem.

Researchers should also evaluate detectors across diverse media formats and realistic environments. Testing on multiple datasets, different compression levels, varying video qualities, and previously unseen manipulation methods can provide a more accurate assessment of real-world reliability.

VII. CONCLUSION

Deepfake technology presents a growing threat to the authenticity and reliability of digital media. This systematic review examined deepfake detection approaches from 2018 to 2025, focusing on deep learning, traditional machine learning, and classical image-processing techniques. The analysis of studies using FaceForensics++, DFDC, and Celeb-DF demonstrates that cross-dataset generalization remains one of the most significant challenges in the field.

Transformer-based methods showed better generalization than the evaluated CNN-based methods, with an observed performance decline of approximately 11.33%, while CNN-based methods experienced declines exceeding 15%. However, Transformer approaches required approximately 3–5 times greater computational resources. Traditional machine learning methods, particularly Random Forest, showed competitive performance in some settings while offering considerably lower computational requirements.

Overall, the findings indicate that successful deepfake detection requires a balance between performance, generalization, computational efficiency, and robustness. Future research should therefore focus on developing practical detection systems capable of identifying unseen deepfake manipulations across diverse media formats and real-world environments rather than optimizing solely for benchmark accuracy.

REFERENCES

- [1] Rössler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies, and M. Nießner, “FaceForensics++: Learning to detect manipulated facial images,” in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, 2019, pp. 1–11.
- [2] H. Kim, P. Garrido, A. Tewari, W. Xu, J. Thies, M. Nießner, P. Pérez, C. Richardt, M. Zollhöfer, and C. Theobalt, “Deep video portraits,” *arXiv preprint arXiv:1805.11714*, 2018.
- [3] X. Zhang, S. Karaman, and S.-F. Chang, “Detecting and simulating artifacts in GAN fake images,” in *Proc. IEEE Int. Workshop Inf. Forensics Security (WIFS)*, 2019.
- [4] J.-W. Seow, M.-K. Lim, R. C.-W. Phan, and J. K. Liu, “A comprehensive overview of deepfake: Generation, detection, datasets, and opportunities,” *Neurocomputing*, 2022.
- [5] H. Dang, F. Liu, J. Stehouwer, X. Liu, and A. K. Jain, “On the detection of digital face manipulation,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2020, pp. 5781–5790.
- [6] Y. Qian, G. Yin, L. Sheng, Z. Chen, and J. Shao, “Thinking in frequency: Face forgery detection by mining frequency-aware clues,” in *Computer Vision—ECCV 2020*, vol. 12357. Cham, Switzerland: Springer, 2020, pp. 86–103, doi: 10.1007/978-3-030-58610-2_6.
- [7] Y. Luo, Y. Zhang, J. Yan, and W. Liu, “Generalizing face forgery detection with high-frequency features,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2021, pp. 16317–16326.
- [8] L. Jiang, R. Li, W. Wu, C. Qian, and C. C. Loy, “DeeperForensics-1.0: A large-scale dataset for real-world face forgery detection,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2020, pp. 2889–2898, doi: 10.1109/CVPR42600.2020.00296.
- [9] D. Afchar, V. Nozick, J. Yamagishi, and I. Echizen, “MesoNet: A compact facial video forgery detection network,” in *Proc. IEEE Int. Workshop Inf. Forensics Security (WIFS)*, 2018, pp. 1–7.
- [10] Y. Li, M.-C. Chang, and S. Lyu, “In Ictu Oculi: Exposing AI generated fake face videos by detecting eye blinking,” in *Proc. IEEE Int. Workshop Inf. Forensics Security (WIFS)*, 2018, pp. 1–7.
- [11] Y. Li, X. Yang, P. Sun, H. Qi, and S. Lyu, “Celeb-DF: A new dataset for DeepFake forensics,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2020, pp. 3207–3215.
- [12] B. Dolhansky, R. Howes, B. Pflaum, N. Baram, and C. C. Ferrer, “The DeepFake Detection Challenge (DFDC) dataset,” *arXiv preprint arXiv:2006.07397*, 2020.
- [13] H. H. Nguyen, F. Fang, J. Yamagishi, and I. Echizen, “Multi-task learning for detecting and segmenting manipulated facial images and videos,” in *Proc. IEEE Int. Conf. Biometrics Theory, Applications and Systems (BTAS)*, 2019, pp. 1–8.